

A chain rule for the expected suprema of Gaussian processes

Andreas Maurer

Adalbertstrasse 55
D-80799 München, Germany
am@andreas-maurer.eu

Abstract. The expected supremum of a Gaussian process indexed by the image of an index set under a function class is bounded in terms of separate properties of the index set and the function class. The bound is relevant to the estimation of nonlinear transformations or the analysis of learning algorithms whenever hypotheses are chosen from composite classes, as is the case for multi-layer models.

1 Introduction

Rademacher and Gaussian averages ([1], see also [5],[11]) provide an elegant method to demonstrate generalization for a wide variety of learning algorithms and are particularly well suited to analyze kernel machines, where the use of more classical methods relying on covering numbers becomes cumbersome.

To briefly describe the use of Gaussian averages (Rademacher averages will not concern us), let $Y \subseteq \mathbb{R}^n$ and let γ be a vector $\gamma = (\gamma_1, \dots, \gamma_n)$ of independent standard normal variables. We define the (expected supremum of the) Gaussian average of Y as

$$G(Y) = \mathbb{E} \sup_{\mathbf{y} \in Y} \langle \gamma, \mathbf{y} \rangle, \quad (1)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in \mathbb{R}^n . Consider a loss class F of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, where \mathcal{X} is some space of examples (such as input-output pairs), a sample $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ of observations and write $F(\mathbf{x})$ for the subset of \mathbb{R}^n given by $F(\mathbf{x}) = \{(f(x_1), \dots, f(x_n)) : f \in F\}$. Then we have the following result [1].

Theorem 1. *Let the members of F take values in $[0, 1]$ and let X, X_1, \dots, X_n be iid random variables with values in \mathcal{X} , $\mathbf{X} = (X_1, \dots, X_n)$. Then for $\delta > 0$ with probability at least $1 - \delta$ we have for every $f \in F$ that*

$$\mathbb{E} f(X) \leq \frac{1}{n} \sum f(X_i) + \frac{\sqrt{2\pi}}{n} G(F(\mathbf{X})) + \sqrt{\frac{9 \ln 2 / \delta}{2n}},$$

where the expectation in the definition (1) of $G(F(\mathbf{X}))$ is conditional to the sample \mathbf{X} .

The utility of Gaussian averages is not limited to functions with values in $[0, 1]$. For real functions ϕ with Lipschitz constant $L(\phi)$ we have $G((\phi \circ F)(\mathbf{x})) \leq L(\phi) G(F(\mathbf{x}))$ (see also Slepian's Lemma, [6], [4]), where $\phi \circ F$ is the class $\{x \mapsto \phi(f(x)) : f \in F\}$.

The inequality $G((\phi \circ F)(\mathbf{x})) \leq L(\phi) G(F(\mathbf{x}))$, which in the above form holds also for Rademacher averages [10], is extremely useful and in part responsible for the success of these complexity measures. For Gaussian averages it holds in a more general sense: if $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ has Lipschitz constant $L(\phi)$ with respect to the Euclidean distances, then $G(\phi(Y)) \leq L(\phi) G(Y)$. This is a direct consequence of Slepian's Lemma and can be applied to the analysis of clustering or learning to learn ([9] and [8]).

But what if we also want some freedom in the choice of ϕ *after* seeing the data? If the class of Lipschitz functions considered has small cardinality, a union bound can be used. If it is very large one can try to use covering numbers, but the matter soon becomes quite complicated and destroys the elegant simplicity of the method.

These considerations lead to a more general question: given a set $Y \subset \mathbb{R}^n$ and a class F of functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, how can we bound the Gaussian average $G(F(Y)) = G(\{f(y) : f \in F, y \in Y\})$ in terms of separate properties of Y and F , properties which should preferably very closely resemble Gaussian averages? If \mathcal{H} is some class of functions mapping samples to \mathbb{R}^n and $Y = \mathcal{H}(\mathbf{x})$, then the bound is on $G(F(Y)) = G((F \circ \mathcal{H})(\mathbf{x}))$, so our question is relevant to the estimation of composite functions in general. Such estimates are necessary for multitask feature-learning, where \mathcal{H} is a class of feature maps and F is vector-valued, with components chosen independently for each task. Other potential applications are to the currently popular subject of deep learning, where we consider functional concatenations as in $\mathcal{F}_M \circ \mathcal{F}_{M-1} \circ \dots \circ \mathcal{F}_1$.

The present paper gives a preliminary answer. To state it we introduce some notation. We will always take $\gamma = (\gamma_1, \dots)$ to be a random vector whose components are independent standard normal variables, while $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denote norm and inner product in a Euclidean space, the dimension of which is determined by context, as is the dimension of the vector γ .

Definition 1. If $Y \subseteq \mathbb{R}^n$ we set

$$D(Y) = \sup_{\mathbf{y}, \mathbf{y}' \in Y} \|\mathbf{y} - \mathbf{y}'\| \text{ and } G(Y) = \mathbb{E} \sup_{\mathbf{y} \in Y} \langle \gamma, \mathbf{y} \rangle.$$

If F is a class of functions $f : Y \rightarrow \mathbb{R}^m$ we set

$$L(F, Y) = \sup_{\mathbf{y}, \mathbf{y}' \in Y, \mathbf{y} \neq \mathbf{y}'} \sup_{f \in F} \frac{\|f(\mathbf{y}) - f(\mathbf{y}')\|}{\|\mathbf{y} - \mathbf{y}'\|} \text{ and}$$

$$R(F, Y) = \sup_{\mathbf{y}, \mathbf{y}' \in Y, \mathbf{y} \neq \mathbf{y}'} \mathbb{E} \sup_{f \in F} \frac{\langle \gamma, f(\mathbf{y}) - f(\mathbf{y}') \rangle}{\|\mathbf{y} - \mathbf{y}'\|}.$$

We also write $F(Y) = \{f(\mathbf{y}) : f \in F, \mathbf{y} \in Y\}$. When there is no ambiguity we write $L(F) = L(F, Y)$ and $R(F) = R(F, Y)$.

Then $D(Y)$ is the diameter of Y , and $G(Y)$ is the Gaussian average already introduced above. $L(F)$ is the smallest Lipschitz constant acceptable for all $f \in F$, and the more unusual quantity $R(F)$ can be viewed as a Gaussian average of Lipschitz quotients. In section 3.1 we give some properties of $R(F)$. Our main result is the following chain rule.

Theorem 2. *Let $Y \subset \mathbb{R}^n$ be finite, F a finite class of functions $f : Y \rightarrow \mathbb{R}^m$. Then there are universal constants C_1 and C_2 such that for any $\mathbf{y}_0 \in Y$*

$$G(F(Y)) \leq C_1 L(F) G(Y) + C_2 D(Y) R(F) + G(F(\mathbf{y}_0)). \quad (2)$$

We make some general remarks on the implications of our result.

1. The requirement of finiteness for Y and F is a simplification to avoid issues of measurability. The cardinality of these sets plays no role.

2. The constants C_1 and C_2 as they result from the proof are rather large, because they accumulate the constants of Talagrand's majorizing measure theorem and generic chaining [6][14][15][16]. This is a major shortcoming and the reason why our result is regarded as preliminary. Is there another proof of a similar result, avoiding majorizing measures and resulting in smaller constants? This question is the subject of current research.

3. The first term on the right hand side of (2) describes the complexity inherited from the bottom layer Y (which we may think of as $\mathcal{H}(\mathbf{x})$), and it depends on the top layer F only through the Lipschitz constant $L(F)$. The other two terms represent the complexity of the top layer, depending on the bottom layer only through the diameter $D(Y)$ of Y . If Y has unit diameter and the functions in F are contractions, then the two layers are completely decoupled in the bound. This decoupling is the most attractive property of our result.

4. Apart from the large constants the inequality is tight in at least two situations: first, if $Y = \{\mathbf{y}_0\}$ is a singleton, then only the last term remains, and we recover the Gaussian average of $\mathcal{F}(\mathbf{y}_0)$. This also shows that the last term cannot be eliminated. On the other hand if F consists of a single Lipschitz function ϕ , then we recover (up to a constant) the inequality $G(\phi(Y)) \leq L(\phi) G(Y)$ above.

5. The bound can be iterated to multiple layers by re-substitution of $F(Y)$ in place of Y . A corresponding formula is given in Section 3, where we also sketch applications to vector-valued function classes.

The next section gives a proof of Theorem 2, then we explain how our result can be applied to machine learning. The last section is devoted to the proof of a technical result encapsulating our use of majorizing measures.

2 Proving the chain rule

To prove Theorem 2 we need the theory of majorizing measures and generic chaining. Our use of these techniques is summarized in the following theorem, which is also the origin of our large constants.

Theorem 3. *Let $X_{\mathbf{y}}$ be a random process indexed by a finite set $Y \subset \mathbb{R}^n$. Suppose that there is a number $K \geq 1$ such that for any distinct members $\mathbf{y}, \mathbf{y}' \in Y$ and any $s > 0$*

$$\Pr \{X_{\mathbf{y}} - X_{\mathbf{y}'} > s\} \leq K \exp \left(\frac{-s^2}{2 \|\mathbf{y} - \mathbf{y}'\|^2} \right) \quad (3)$$

Then for any $\mathbf{y}_0 \in Y$

$$\mathbb{E} \left[\sup_{\mathbf{y} \in Y} X_{\mathbf{y}} - X_{\mathbf{y}_0} \right] \leq C' G(Y) + C'' D(Y) \sqrt{\ln K},$$

where C' and C'' are universal constants.

This is obtained from Talagrand's majorizing measure theorem (Theorem 6 below) combined with generic chaining [16]. An early version of a similar result is Theorem 15 in [13], where the author remarks that his method of proof (which we also use) is very indirect, and that a more direct proof would be desirable. In Section 4 we do supply a proof, largely because the dependence on K , which can often be swept under the carpet, plays a crucial role in our arguments below.

We also need the following Gaussian concentration inequality (Tsirelson-Ibragimov-Sudakov inequality, Theorem 5.6 in [4]).

Theorem 4. *Let $F : \mathbb{R}^n \rightarrow \mathbb{R}$ be L -Lipschitz. Then for any $s > 0$*

$$\Pr \{F(\boldsymbol{\gamma}) > \mathbb{E}F(\boldsymbol{\gamma}) + s\} \leq e^{-s^2/(2L^2)}.$$

To conclude the preparation for the proof of Theorem 2 we give a simple lemma.

Lemma 1. *Suppose a random variable X satisfies $\Pr \{X - A > s\} \leq e^{-s^2}$, for any $s > 0$. Then*

$$\forall s > 0, \Pr \{X > s\} \leq e^{A^2} e^{-s^2/2}.$$

Proof. For $s \leq A$ the conclusion is trivial, so suppose that $s > A$. From $s^2 = (s - A + A)^2 \leq 2(s - A)^2 + 2A^2$ we get $(s - A)^2 \geq (s^2/2) - A^2$, so

$$\Pr \{X > s\} = \Pr \{X - A > s - A\} \leq e^{-(s-A)^2} \leq e^{A^2} e^{-s^2/2}.$$

■

Proof (of Theorem 2). The result is trivial if F consists only of constants, so we can assume that $L(F) > 0$. For $\mathbf{y}, \mathbf{y}' \in Y$ define a function $F : \mathbb{R}^m \rightarrow \mathbb{R}$ by

$$F(\mathbf{z}) = \sup_{f \in F} \langle \mathbf{z}, f(\mathbf{y}) - f(\mathbf{y}') \rangle.$$

F is Lipschitz with Lipschitz constant bounded by $\sup_{f \in F} \|f(\mathbf{y}) - f(\mathbf{y}')\| \leq L(F) \|\mathbf{y} - \mathbf{y}'\|$. Writing $Z_{\mathbf{y}, \mathbf{y}'} = F(\gamma)$, it then follows from Gaussian concentration (Theorem 4) that

$$\Pr \{Z_{\mathbf{y}, \mathbf{y}'} > \mathbb{E}Z_{\mathbf{y}, \mathbf{y}'} + s\} \leq \exp \left(\frac{-s^2}{2L(F)^2 \|\mathbf{y} - \mathbf{y}'\|^2} \right).$$

Since by definition $\mathbb{E}Z_{\mathbf{y}, \mathbf{y}'} \leq R(F) \|\mathbf{y} - \mathbf{y}'\|$, Lemma 1 gives

$$\Pr \{Z_{\mathbf{y}, \mathbf{y}'} > s\} \leq \exp \left(\frac{R(F)^2}{2L(F)^2} \right) \exp \left(\frac{-s^2}{4L(F)^2 \|\mathbf{y} - \mathbf{y}'\|^2} \right).$$

Now define a process $X_{\mathbf{y}}$, indexed by Y , as

$$X_{\mathbf{y}} = \frac{1}{\sqrt{2}L(F)} \sup_{f \in F} \langle \gamma, f(\mathbf{y}) \rangle.$$

Since $X_{\mathbf{y}} - X_{\mathbf{y}'} \leq Z_{\mathbf{y}, \mathbf{y}'} / (\sqrt{2}L(F))$ we have

$$\begin{aligned} \Pr \{X_{\mathbf{y}} - X_{\mathbf{y}'} > s\} &\leq \Pr \{Z_{\mathbf{y}, \mathbf{y}'} > \sqrt{2}L(F)s\} \\ &\leq \exp \left(\frac{R(F)^2}{2L(F)^2} \right) \exp \left(\frac{-s^2}{2\|\mathbf{y} - \mathbf{y}'\|^2} \right) \end{aligned}$$

and by Theorem 3, with $K = \exp \left(R(F)^2 / (2L(F)^2) \right) \geq 1$,

$$\mathbb{E} \sup_{\mathbf{y} \in Y} (X_{\mathbf{y}} - X_{\mathbf{y}_0}) \leq C'G(Y) + C''D(Y) \frac{R(F)}{\sqrt{2}L(F)}.$$

Multiplication by $\sqrt{2}L(F)$ then gives

$$\mathbb{E} \sup_{\mathbf{y} \in Y} \left(\sup_{f \in F} \langle \gamma, f(\mathbf{y}) \rangle - \sup_{f \in F} \langle \gamma, f(\mathbf{y}_0) \rangle \right) \leq C_1 L(F) G(Y) + C_2 D(Y) R(F)$$

with $C_1 = \sqrt{2}C'$ and $C_2 = C''$. ■

3 Applications

We first give some elementary properties of the quantity $R(F, Y)$ which appears in Theorem 2. Then we apply Theorem 2 to a two layer kernel machine and give a bound for multi-task learning of low-dimensional representations.

3.1 Some properties of $R(F)$

Recall the definition of $R(F, Y)$. If $Y \subseteq \mathbb{R}^n$ and F consists of functions $f : Y \rightarrow \mathbb{R}^m$

$$R(F, Y) = \sup_{\mathbf{y}, \mathbf{y}' \in Y, \mathbf{y} \neq \mathbf{y}'} \mathbb{E} \sup_{f \in F} \frac{\langle \gamma, f(\mathbf{y}) - f(\mathbf{y}') \rangle}{\|\mathbf{y} - \mathbf{y}'\|}.$$

$R(F)$ is itself a supremum of Gaussian averages. For $\mathbf{y}, \mathbf{y}' \in Y$ let $\Delta F(\mathbf{y}, \mathbf{y}') \subseteq \mathbb{R}^m$ be the set of quotients

$$\Delta F(\mathbf{y}, \mathbf{y}') = \left\{ \frac{f(\mathbf{y}) - f(\mathbf{y}')}{\|\mathbf{y} - \mathbf{y}'\|} : f \in F \right\}.$$

It follows from the definition that $R(F, Y) = \sup_{\mathbf{y}, \mathbf{y}' \in Y, \mathbf{y} \neq \mathbf{y}'} G(\Delta F(\mathbf{y}, \mathbf{y}'))$. We record some simple properties. Recall that for a set S in a real vector space the convex hull $Co(S)$ is defined as

$$Co(S) = \left\{ \sum_{i=1}^n \alpha_i z_i : n \in \mathbb{N}, z_i \in S, \alpha_i \geq 0, \sum_i \alpha_i = 1 \right\}.$$

Theorem 5. *Let $Y \subseteq \mathbb{R}^n$ and let F and \mathcal{H} be classes of functions $f : Y \rightarrow \mathbb{R}^m$. Then*

- (i) *If $F \subseteq \mathcal{H}$ then $R(F, Y) \leq R(\mathcal{H}, Y)$.*
- (ii) *If $Y \subseteq Y'$ then $R(F, Y) \leq R(F, Y')$.*
- (iii) *If $c \geq 0$ then $R(cF, Y) = cR(F, Y)$.*
- (iv) *$R(F + \mathcal{H}, Y) \leq R(F, Y) + R(\mathcal{H}, Y)$.*
- (v) *$R(F, Y) = R(Co(F), Y)$.*
- (vi) *If $Z \subseteq \mathbb{R}^K$ and $\phi : Z \rightarrow \mathbb{R}^n$ has Lipschitz constant $L(\phi)$ and the members of F are defined on $\phi(Z)$, then $R(F \circ \phi, Z) \leq L(\phi) R(F, \phi(Z))$.*
- (vii) *$R(F) \leq L(F) \sqrt{2 \ln |F|}$.*

Remarks:

1. From (ii) we get $R(F, Y) \leq R(F, \mathbb{R}^n)$. In applications where $Y = \mathcal{H}(\mathbf{x})$ the quantity $R(F, \mathcal{H}(\mathbf{x}))$ is data-dependent, but $R(F, \mathbb{R}^n)$ is sometimes easier to bound.

2. We see that the properties of $R(F)$ largely parallel the properties of the Gaussian averages themselves, except for the inequality $G(\phi(Y)) \leq L(\phi) G(Y)$, for which there doesn't seem to be an analogous property of $R(F)$. Instead we have a 'backwards' version of it with (vi) above, with a rather trivial proof below.

3. Of course (vii) is relevant only when $\ln |F|$ is reasonably small and serves the comparison of Theorem 2 to alternative bounds.

Proof. (i)-(iii) are obvious from the definition. (iv) follows from linearity of the inner product and the triangle inequality for the supremum. To see (v) first note that $R(F) \leq R(Co(F))$ follows from (i), while the reverse inequality follows

from

$$\begin{aligned}
& \sup_{\alpha_i \geq 0, \sum \alpha_i = 1} \sup_{f_1, f_2, \dots \in F} \left\langle \gamma, \sum_i \alpha_i f_i(\mathbf{y}) - \sum_i \alpha_i f_i(\mathbf{y}') \right\rangle \\
&= \sup_{\alpha_i \geq 0, \sum \alpha_i = 1} \sup_{f_1, f_2, \dots \in F} \sum_i \alpha_i \langle \gamma, f_i(\mathbf{y}) - f_i(\mathbf{y}') \rangle \\
&\leq \sup_{\alpha_i \geq 0, \sum \alpha_i = 1} \sum_i \alpha_i \sup_{f \in F} \langle \gamma, f(\mathbf{y}) - f(\mathbf{y}') \rangle \\
&= \sup_{f \in F} \langle \gamma, f(\mathbf{y}) - f(\mathbf{y}') \rangle.
\end{aligned}$$

For (vi) we may chose \mathbf{y} and \mathbf{y}' such that $\phi(\mathbf{y}) \neq \phi(\mathbf{y}')$, since otherwise both sides of the inequality to be proved are zero. But then

$$\begin{aligned}
\mathbb{E} \sup_{f \in F \circ \phi} \frac{\langle \gamma, f(\mathbf{y}) - f(\mathbf{y}') \rangle}{\|\mathbf{y} - \mathbf{y}'\|} &= \frac{\|\phi(\mathbf{y}) - \phi(\mathbf{y}')\|}{\|\mathbf{y} - \mathbf{y}'\|} \mathbb{E} \sup_{f \in F} \frac{\langle \gamma, f(\phi(\mathbf{y})) - f(\phi(\mathbf{y}')) \rangle}{\|\phi(\mathbf{y}) - \phi(\mathbf{y}')\|} \\
&\leq L(\phi) \mathbb{E} \sup_{f \in F} \frac{\langle \gamma, f(\phi(\mathbf{y})) - f(\phi(\mathbf{y}')) \rangle}{\|\phi(\mathbf{y}) - \phi(\mathbf{y}')\|}.
\end{aligned}$$

To see (vii) note that for every \mathbf{y} and \mathbf{y}' and every $f \in F$ it follows from Gaussian concentration (Theorem 4) that

$$\Pr \left\{ \frac{\langle \gamma, f(\mathbf{y}) - f(\mathbf{y}') \rangle}{\|\mathbf{y} - \mathbf{y}'\|} > s \right\} \leq e^{-s^2/2L^2}.$$

The conclusion then follows from standard estimates (e.g. [4], section 2.5). ■

3.2 A double layer kernel machine

We use the chain rule to bound the complexity of a double-layer kernel machine. The corresponding optimization problem is clearly non-convex and we are not aware of an efficient optimization method. The model is chosen to illustrate the application of Theorem 2. It is defined as follows.

Assume the data to lie in \mathbb{R}^{m_0} and fix two real numbers Δ_1 and B_1 . On $\mathbb{R}^{m_0} \times \mathbb{R}^{m_0}$ define a (Gaussian radial-basis-function) kernel κ by

$$\kappa(z, z') = \exp \left(\frac{-\|z - z'\|^2}{2\Delta_1^2} \right), \quad z, z' \in \mathbb{R}^{m_0},$$

and let $\phi : \mathbb{R}^{m_0} \rightarrow H$ be the associated feature map, where H is the associated RKHS with inner product $\langle \cdot, \cdot \rangle_H$ and norm $\|\cdot\|_H$ (for kernel methods see . Now we let \mathcal{H} be the class of vector valued functions $h : \mathbb{R}^{m_0} \rightarrow \mathbb{R}^{m_1}$ defined by

$$\mathcal{H} = \left\{ z \in \mathbb{R}^{m_0} \mapsto (\langle w_1, \phi(z) \rangle_H, \dots, \langle w_{m_1}, \phi(z) \rangle_H) : \sum_k \|w_k\|_H^2 \leq B_1^2 \right\}.$$

This can also be written as $\mathcal{H} = \{z \in \mathbb{R}^{m_0} \mapsto W\phi(z) : \|W\|_{HS} \leq B_1\}$, where $\|W\|_{HS}$ is the Hilbert-Schmidt norm of an operator $W : H \rightarrow \mathbb{R}^{m_1}$.

For the function class \mathcal{F} , which we wish to compose with \mathcal{H} , we proceed in a similar way, defining an analogous kernel of width Δ_2 on $\mathbb{R}^{m_1} \times \mathbb{R}^{m_1}$, a corresponding feature map $\psi : \mathbb{R}^{m_1} \rightarrow H$ and a class of real valued functions

$$\mathcal{F} = \{z \in \mathbb{R}^{m_1} \mapsto \langle v, \psi(z) \rangle_H : \|v\|_H \leq B_2\}.$$

We now want high probability bounds on the estimation error for functional compositions $f \circ h$, uniform over $F \circ \mathcal{H}$. To apply our result we should really restrict to finite subsets of F and \mathcal{H} a requirement which we simply ignore. In machine learning we could of course always restrict all representations to some fixed, very high but finite precision.

Fix a sample $\mathbf{x} \in \mathbb{R}^{nm_0}$. Then $Y = \mathcal{H}(\mathbf{x}) \subset \mathbb{R}^{nm_1}$. To use Theorem 2 we define a class F' of functions from \mathbb{R}^{nm_1} to \mathbb{R}^n by

$$F' = \{(y_1, \dots, y_n) \in \mathbb{R}^{nm_1} \mapsto (f(y_1), \dots, f(y_n)) \in \mathbb{R}^n : f \in F\}.$$

Since the first feature map ϕ maps to the unit sphere of H we have

$$\begin{aligned} D(\mathcal{H}(\mathbf{x})) &\leq 2B_1\sqrt{n} \text{ and} \\ G(\mathcal{H}(\mathbf{x})) &= \mathbb{E} \sup_W \sum_{ik} \gamma_{ik} \langle w_k, \phi(x_i) \rangle_H \leq B_1\sqrt{nm_1}. \end{aligned}$$

The feature map corresponding to the Gaussian kernel Δ_2 has Lipschitz constant Δ_2^{-1} . For $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{nm_1}$ we obtain

$$\begin{aligned} \sup_v \left(\sum_i (\langle v, \phi(y_i) \rangle_H - \langle v, \phi(y'_i) \rangle_H)^2 \right)^{1/2} &\leq B_2 \left(\sum_i \|\phi(y_i) - \phi(y'_i)\|_H^2 \right)^{1/2} \\ &\leq B_2 \Delta_2^{-1} \|\mathbf{y} - \mathbf{y}'\|, \end{aligned}$$

so we have $L(F', \mathbb{R}^{nm_1}) \leq B_2 \Delta_2^{-1}$.

On the other hand

$$\begin{aligned} \mathbb{E} \sup_v \sum_i \gamma_i (\langle v, \phi(y_i) \rangle_H - \langle v, \phi(y'_i) \rangle_H) &\leq B_2 \mathbb{E} \left\| \sum_{i=1}^n \gamma_i (\phi(\mathbf{y}_i) - \phi(\mathbf{y}'_i)) \right\| \\ &\leq B_2 \left(\sum_i \|\phi(y_i) - \phi(y'_i)\|_H^2 \right)^{1/2} \\ &\leq B_2 \Delta_2^{-1} \|\mathbf{y} - \mathbf{y}'\|, \end{aligned}$$

so we have $R(F', \mathbb{R}^{nm_1}) \leq B_2 \Delta_2^{-1}$. Furthermore

$$G(F'(h_0(\mathbf{x}))) \leq B_2 \sqrt{n},$$

similar to the bound for $G(\mathcal{H}(\mathbf{x}))$.

For the composite network Theorem 2 gives us the bound

$$G(F'(\mathcal{H}(\mathbf{x}))) \leq C_1 B_1 B_2 \Delta_2^{-1} \sqrt{nm_1} + 2C_2 B_1 B_2 \sqrt{n} \Delta_2^{-1} + B_2 \sqrt{n}.$$

Dividing by n and appealing to Theorem 1 one obtains the uniform bound: with probability at least $1 - \delta$ we have for every $h \in \mathcal{H}$ and every $f \in F$ that

$$\begin{aligned} \mathbb{E}f(h(X)) &\leq \frac{1}{n} \sum f(h(X_i)) + \\ &\quad + \sqrt{\frac{2\pi}{n}} B_2 (B_1 \Delta_2^{-1} (C_1 \sqrt{m_1} + 2C_2) + 1) + \sqrt{\frac{9 \ln 2/\delta}{2n}}. \end{aligned}$$

Remarks.

1. One might object that the result depends heavily on the intermediate dimension m_1 so that only a very classical relationship between sample size and dimension is obtained. In this sense our result only works for intermediate representations of rather low dimension. The mapping stages of \mathcal{H} and F however include nonlinear maps to infinite dimensional spaces.
2. Clearly the above choice of the Gaussian kernel is arbitrary. Any positive semidefinite kernel can be used for the first mapping stage, and the application of the chain rule requires only the Lipschitz property for the second kernel in the definition of F . The Gaussian kernel was only chosen for definiteness.
3. Similarly the choice of the Hilbert-Schmidt norm as a regularizer for W in the first mapping stage is arbitrary, one could equally use another matrix norm. This would result in different bounds for $G(\mathcal{H}(\mathbf{x}))$ and $D(\mathcal{H}(\mathbf{x}))$, incurring a different dependency of our bound on m_1 .

3.3 Multitask learning

As a second illustration we modify the above model to accommodate multitask learning [2][3]. Here one observes a $T \times n$ sample $\mathbf{x} = (x_{ti} : 1 \leq t \leq T, 1 \leq i \leq n) \in \mathcal{X}^{nT}$, where $(x_{ti} : 1 \leq i \leq n)$ is the sample observed for the t -th task. We consider a two layer situation where the bottom-layer \mathcal{H} consists of functions $h : \mathcal{X} \rightarrow \mathbb{R}^m$, and the top layer function class is of the form

$$F^T = \{x \in \mathbb{R}^{m_1} \mapsto \mathbf{f}(x) = (f_1(x), \dots, f_T(x)) \in \mathbb{R}^T : f_t \in F\},$$

where F is some class of functions mapping \mathbb{R}^{m_1} to \mathbb{R} . The functions (or representations) of the bottom layer \mathcal{H} are optimized for the entire sample, in the top layer each function f_t is optimized for the represented data corresponding to the t -th task. In an approach of empirical risk minimization one selects the composed function $\hat{\mathbf{f}} \circ \hat{h}$ which minimizes the task-averaged empirical loss

$$\min_{\mathbf{f} \in F^n, h \in \mathcal{H}} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T f_t(h(x_{it})).$$

We wish to give a general explanation of the potential benefits of this method over the separate learning of functions from $F \circ \mathcal{H}$, as studied in the previous

section. Clearly we must assume that the tasks are related in the sense that the above minimum is small, so any possible benefit can only be a benefit of improved estimation.

For the multitask model a result analogous to Theorem 1 is easily obtained (see e.g. [7]). Let $\mathbf{X} = (X_{ti})$ be a vector of independent random variables with values in \mathcal{X} , where X_{ti} is iid to X_{tj} for all ijt , and let X_t be iid to X_{ti} . Then with probability at least $1 - \delta$ we have for every $\mathbf{f} \in F^n$ and every $h \in \mathcal{H}$

$$\frac{1}{T} \sum_t \mathbb{E} f_t(h(X_t)) \leq \frac{1}{nT} \sum_{ti} f_t(h(X_{ti})) + \frac{\sqrt{2\pi}}{nT} G(F^T \circ \mathcal{H}(\mathbf{X})) + \sqrt{\frac{9 \ln 2/\delta}{2nT}}.$$

Here the left hand side is interpreted as the task averaged risk and

$$G(F^T \circ \mathcal{H}(\mathbf{x})) = \mathbb{E} \sup_{\mathbf{f} \in F^T, h \in \mathcal{H}} \sum_{ti} \gamma_{ti} f_t(h(x_{ti})).$$

For a definite example we take \mathcal{H} and F as in the previous section and observe that now there is an additional factor T on the sample size. This implies the modified bounds $G(\mathcal{H}(\mathbf{x})) \leq B_1 \sqrt{Tnm_1}$ and $D(\mathcal{H}(\mathbf{x})) \leq 2B_1 \sqrt{Tn}$. Also for $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{Tnm_1}$ with $y_{ti}, y'_{ti} \in \mathbb{R}^{m_1}$

$$\begin{aligned} \sup_{\mathbf{f} \in F^T} \sum_{ti} (f_t(y_{ti}) - f_t(y'_{ti}))^2 &\leq \sum_t \sup_{f \in F} \sum_i (f_t(y_{ti}) - f_t(y'_{ti}))^2 \\ &\leq L^2(F, \mathbb{R}^{nm_1}) \sum_t \sum_i \|y_{ti} - y'_{ti}\|^2, \end{aligned}$$

so

$$L(F^T, \mathbb{R}^{Tnm_1}) = L(F, \mathbb{R}^{nm_1}). \quad (4)$$

Therefore $L(F^T, \mathbb{R}^{Tnm_1}) \leq B_2 \Delta_2^{-1}$. Similarly

$$\begin{aligned} &\mathbb{E} \sup_{\mathbf{f} \in F^T} \sum_{ti} \gamma_{ti} (f_t(y_{ti}) - f_t(y'_{ti})) \\ &= \sum_t \mathbb{E} \sup_{f \in F} \sum_i \gamma_{ti} (f_t(y_{ti}) - f_t(y'_{ti})) \\ &\leq \sqrt{T} \left(\sum_t \left(\mathbb{E} \sup_{f \in F} \sum_i \gamma_{ti} (f_t(y_{ti}) - f_t(y'_{ti})) \right)^2 \right)^{1/2} \\ &\leq \sqrt{T} \left(\sum_t R^2(F, \mathbb{R}^{nm_1}) \sum_i \|y_{ti} - y'_{ti}\|^2 \right)^{1/2} \\ &= \sqrt{T} R(F, \mathbb{R}^{nm_1}) \|\mathbf{y} - \mathbf{y}'\|. \end{aligned}$$

We conclude that

$$R(F^T, \mathbb{R}^{nmT}) \leq \sqrt{T} R(F, \mathbb{R}^{nm}), \quad (5)$$

in the given case

$$R(F^T, \mathbb{R}^{nmT}) \leq \sqrt{T} B_2 \Delta_2^{-1}.$$

Also

$$\begin{aligned}
G(F^T(h_0(\mathbf{x}))) &= \mathbb{E} \sup_{\mathbf{f} \in F^T} \sum_{ti} \gamma_{ti} f_t(h_0(x_{ti})) \\
&= \sum_t \mathbb{E} \sup_{f \in F} \sum_i \gamma_{ti} f(h_0(x_{ti})) \\
&\leq TG(F(h_0(\mathbf{x}))), \tag{6}
\end{aligned}$$

so that here $G(F^T(h_0(\mathbf{x}))) \leq B_2 T \sqrt{n}$. The chain rule then gives

$$G(F \circ \mathcal{H}(\mathbf{x})) \leq C_1 B_1 B_2 \Delta_2^{-1} \sqrt{T n m_1} + (2C_2 B_1 \Delta_2^{-1} + 1) B_2 T \sqrt{n},$$

where the first term represents the complexity of \mathcal{H} and the second that of F^T . Dividing by nT we obtain as the dominant term for the estimation error

$$C_1 B_1 B_2 \Delta_2^{-1} \sqrt{\frac{m_1}{nT}} + \frac{(2C_2 B_1 \Delta_2^{-1} + 1) B_2}{\sqrt{n}}.$$

This reproduces a general property of multitask learning [3]: in the limit $T \rightarrow \infty$ the contribution of the common representation (including the intermediate dimension m_1) to the estimation error vanishes. There remains only the cost of estimating the task specific functions in the top layer.

We have obtained this result for a very specific model. The relations (4), (5) and (6) for $L(F^T)$, $R(F^T)$ and $G(F^T(h_0(\mathbf{x})))$ are nevertheless independent of the exact model, so the chain rule could be made the basis of a fairly general result about multitask feature learning.

3.4 Iteration of the bound

We apply the chain rule to multi-layered or "deep" learning machines, a subject which appears to be of some current interest. Here we have function classes F_1, \dots, F_K , where F_k consists of functions $f: \mathbb{R}^{n_{k-1}} \rightarrow \mathbb{R}^{n_k}$ and we are interested in the generalization properties of the composite class

$$F_K \circ \dots \circ F_1 = \{\mathbf{x} \in \mathbb{R}^{n_0} \mapsto f_K(f_{K-1}(\dots(f_1(\mathbf{x})))) : f_k \in F_k\}.$$

To state our bound we are given some sample \mathbf{x} in \mathbb{R}^{n_0} and introduce the notation

$$\begin{aligned}
Y_0 &= \mathbf{x} \\
Y_k &= F_k(Y_{k-1}) = F_k \circ \dots \circ F_1(\mathbf{x}) \subseteq \mathbb{R}^{n_k}, \text{ for } k > 0 \\
G_k &= \min_{\mathbf{y} \in Y_{k-1}(\mathbf{x})} G(F_k(\mathbf{y})).
\end{aligned}$$

Under the convention that the product over an empty index set is 1, induction shows that

$$G(Y_K) \leq \sum_{k=1}^K \left(C_1^{K-k} \prod_{j=k+1}^K L(F_j) \right) (C_2 D(Y_{k-1}) R(F_k) + G_k).$$

Clearly the large constants are prohibitive for any useful quantitative prediction of generalization, but qualitative statements are possible. Observe for example that, apart from C_1 and the Lipschitz constants, each layer only makes an additive contribution to the overall complexity. More specifically, for machine learning with a sample of size n , we can make the assumptions $n_k = nm_k$, where m_k is the dimension of the k -th intermediate representations, and it is reasonable to postulate $\max \{G_k, D(Y_k) R(F_k)\} \leq Cn^p$, where C is some constant not depending on n and p is some exponent $p < 1$ (for multi-layered kernel machines with Lipschitz feature maps we would have $p = 1/2$ - see above). Then the above expression is of order n^p and Theorem 1 yields a uniform law of large numbers for the multi-layered class, with a uniform bound on the estimation error decreasing as n^{p-1} .

4 Proof of Theorem 3

Talagrand has proved the following result ([14]).

Theorem 6. *There are universal constants $r \geq 2$ and C such that for every finite $Y \subset \mathbb{R}^n$ there is an increasing sequence of partitions \mathcal{A}_k of Y and a probability measure μ on Y , such that, whenever $A \in \mathcal{A}_k$ then $D(A) \leq 2r^{-k}$ and*

$$\sup_{\mathbf{y} \in Y} \sum_{k > k_0}^{\infty} r^{-k} \sqrt{\ln \frac{1}{\mu(A_k(\mathbf{y}))}} \leq C G(Y),$$

where $A_k(\mathbf{y})$ denotes the unique member of \mathcal{A}_k which contains \mathbf{y} , and k_0 is the largest integer k satisfying

$$2r^{-k} \geq D(Y) = \sup_{\mathbf{y}, \mathbf{y}' \in Y} \|\mathbf{y} - \mathbf{y}'\|$$

Observe that $2r^{-k_0} \geq D(Y)$, so we can assume $\mathcal{A}_{k_0} = \{Y\}$. As explained in [14], the above Theorem is equivalent to the existence of a measure μ on Y such that

$$\sup_{\mathbf{y} \in Y} \int_0^{\infty} \sqrt{\ln \frac{1}{\mu(B(\mathbf{y}, \epsilon))}} d\epsilon \leq C G(Y),$$

where C is some other universal constant and $B(\mathbf{y}, \epsilon)$ is the ball of radius ϵ centered at \mathbf{y} . The latter is perhaps the more usual formulation of the majorizing measure theorem.

We will use Talagrand's theorem to prove Theorem 3, but before please note the inequality

$$D(Y) \leq \sqrt{2\pi} G(Y), \quad (7)$$

which follows from

$$\begin{aligned} \sup_{\mathbf{y}, \mathbf{y}' \in Y} \|\mathbf{y} - \mathbf{y}'\| &= \sqrt{\frac{\pi}{2}} \sup_{\mathbf{y}, \mathbf{y}' \in Y} \mathbb{E} |\langle \gamma, \mathbf{y} - \mathbf{y}' \rangle| \\ &\leq \sqrt{\frac{\pi}{2}} \mathbb{E} \sup_{\mathbf{y}, \mathbf{y}' \in Y} |\langle \gamma, \mathbf{y} - \mathbf{y}' \rangle| = \sqrt{\frac{\pi}{2}} \mathbb{E} \sup_{\mathbf{y}, \mathbf{y}' \in Y} \langle \gamma, \mathbf{y} - \mathbf{y}' \rangle. \end{aligned}$$

In the first equality we used the fact that $\|v\| = \sqrt{\pi/2\mathbb{E}} |\langle \gamma, v \rangle|$ for any vector v .

Proof (of Theorem 3.). Let μ and \mathcal{A}_k be as determined for Y by Theorem 6. First we claim that for any $\delta \in (0, 1)$

$$\Pr \left\{ \exists \mathbf{y} \in Y : X_{\mathbf{y}} - X_{\mathbf{y}_0} > \sum_{k > k_0} r^{-k+1} \sqrt{8 \ln \left(\frac{2^{k-k_0} K}{\mu(A(\mathbf{y})) \delta} \right)} \right\} < \delta. \quad (8)$$

For every $k > k_0$ and every $A \in \mathcal{A}_k$ let $\pi(A)$ be some element chosen from A . We set $\pi(Y) = \mathbf{y}_0$. We denote $\pi_k(\mathbf{y}) = \pi(A_k(\mathbf{y}))$. This implies the chaining identity:

$$X_{\mathbf{y}} - X_{\mathbf{y}_0} = \sum_{k > k_0} (X_{\pi_k(\mathbf{y})} - X_{\pi_{k-1}(\mathbf{y})}).$$

For $k > k_0$ and $A \in \mathcal{A}_k$ use \hat{A} to denote the unique member of \mathcal{A}_{k-1} such that $A \subseteq \hat{A}$. Since for $A \in \mathcal{A}_k$ both $\pi(A)$ and $\pi(\hat{A})$ are members of $\hat{A} \in \mathcal{A}_{k-1}$ we must have $\|\pi(A) - \pi(\hat{A})\| \leq 2r^{-k+1}$. Also note $\pi_{k-1}(\mathbf{y}) = \pi(\hat{A}_k(\mathbf{y})) = \pi((A_k(\pi_k(\mathbf{y})))^\wedge)$. For $k \geq k_0$ we define a function $\xi_k : \mathcal{A}_k \rightarrow \mathbb{R}_+$ as follows:

$$\xi_k(A) = r^{-k+1} \sqrt{8 \ln \left(\frac{2^{k-k_0} K}{\mu(A) \delta} \right)}.$$

To prove the claim we have to show that

$$\Pr \left\{ \exists \mathbf{y} \in Y : X_{\mathbf{y}} - X_{\mathbf{y}_0} - \sum_{k > k_0} \xi_k(A_k(\mathbf{y})) > 0 \right\} < \delta.$$

Denote the left hand side of this inequality with P . By the chaining identity

$$P \leq \Pr \left\{ \exists \mathbf{y} : \sum_{k > k_0} (X_{\pi_k(\mathbf{y})} - X_{\pi_{k-1}(\mathbf{y})} - \xi_k(A_k(\mathbf{y}))) > 0 \right\}.$$

If the sum is positive, at least one of the terms has to be positive, so

$$P \leq \Pr \{ \exists \mathbf{y}, k > k_0 : (X_{\pi_k(\mathbf{y})} - X_{\pi_{k-1}(\mathbf{y})} - \xi_k(A_k(\mathbf{y}))) > 0 \}.$$

The event on the right hand side can also be written as

$$\left\{ \exists k > k_0, \exists A \in \mathcal{A}_k : X_{\pi(A)} - X_{\pi(\hat{A})} > \xi_k(A) \right\},$$

and a union bound gives

$$\begin{aligned}
P &\leq \sum_{k>k_0} \sum_{A \in \mathcal{A}_k} \Pr \left\{ X_{\pi(A)} - X_{\pi(\hat{A})} > \xi_k(A) \right\} \\
&\leq \sum_{k>k_0} \sum_{A \in \mathcal{A}_k} K \exp \left(\frac{-\xi_k(A)^2}{2 \left\| \pi(A) - \pi(\hat{A}) \right\|^2} \right) \\
&\leq \sum_{k>k_0} \sum_{A \in \mathcal{A}_k} K \exp \left(\frac{-\xi_k(A)^2}{2 (2^{r-k+1})^2} \right),
\end{aligned}$$

where we used the bound (3) in the second and the bound on $\left\| \pi(A) - \pi(\hat{A}) \right\|$ in the third inequality. Using the definition of $\xi_k(A)$ the last expression is equal to

$$\delta \sum_{k>k_0} \frac{1}{2^{k-k_0}} \sum_{A \in \mathcal{A}_k} \mu(A) = \delta \sum_{k>k_0} \frac{1}{2^{k-k_0}} = \delta,$$

because μ is a probability measure. This establishes the claim.

Now, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, with probability at least $1 - \delta$

$$\begin{aligned}
\sup_{\mathbf{y}} X_{\mathbf{y}} - X_{\mathbf{y}_0} &\leq r \sum_{k>k_0} r^{-k} \sqrt{8 \ln \left(\frac{1}{\mu(A_k(\mathbf{y}))} \right)} \\
&\quad + r^{-k_0+1} \sum_{k>0} r^{-k+1} \sqrt{8 \ln \left(\frac{2^k K}{\delta} \right)} \\
&\leq \sqrt{8} r C G(Y) + \sqrt{8} r^{-k_0+1} \sum_{k>0} \sqrt{k} r^{-k+1} \sqrt{\ln \left(\frac{2K}{\delta} \right)},
\end{aligned}$$

where we used Talagrand's theorem and the fact that $K > 1$. By the definition of k_0 we have $r^{-k_0+1} \leq r^2 D(Y)/2$, so this is bounded by

$$C''' G(Y) + C'''' D(Y) \sqrt{\ln \left(\frac{2K}{\delta} \right)},$$

with $C''' = \sqrt{8} r C$ and $C'''' = \sqrt{8} (r^2/2) \sum_{k>0} \sqrt{k} r^{-k+1}$. Converting the last bound into a tail bound and integrating we obtain

$$\begin{aligned}
\mathbb{E} \left[\sup_{\mathbf{y}} X_{\mathbf{y}} - X_{\mathbf{y}_0} \right] &\leq C''' G(Y) + C'''' D(Y) \left(\sqrt{\ln 2K} + \frac{\sqrt{\pi}}{2} \right) \\
&\leq C''' G(Y) + 3C'''' D(Y) \sqrt{\ln 2K} \\
&\leq \left(C''' + 3\sqrt{2\pi \ln 2} C'''' \right) G(Y) + 3C'''' D(Y) \sqrt{\ln K},
\end{aligned}$$

where we again used $K \geq 1$ in the second inequality and (7) in the last inequality. This gives the conclusion with $C' = C''' + 3\sqrt{2\pi \ln 2} C''''$ and $C'' = 3C''''$. ■

References

1. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463–482, 2002.
2. J. Baxter, Theoretical Models of Learning to Learn, in *Learning to Learn*, S. Thrun, L. Pratt Eds. Springer 1998
3. J. Baxter, A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research* 12:149–198, 2000.
4. S. Boucheron, G. Lugosi, P. Massart. *Concentration Inequalities*, Oxford University Press, 2013
5. V. I. Koltchinskii and D. Panchenko. *Rademacher processes and bounding the risk of function learning*. In E. Gine, D. Mason, and J. Wellner, editors, *High Dimensional Probability II*, pages 443–459. 2000.
6. M. Ledoux, M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.
7. A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.
8. A. Maurer. *Transfer bounds for linear feature learning*. Machine Learning, 75(3): 327–350, 2009.
9. A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11): 5839–5846, 2010.
10. R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4: 839–860, 2003.
11. S. Mendelson. l -norm and its application to learning theory. *Positivity*, 5:177–191, 2001.
12. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
13. M. Talagrand. Regularity of Gaussian processes. *Acta Mathematica*. 159: 99–149, 1987.
14. M. Talagrand. A simple proof of the majorizing measure theorem. *Geometric and Functional Analysis*. Vol 2, No.1: 118–125, 1992.
15. M. Talagrand. Majorizing measures without measures. *Ann. Probab.* 29: 411–417, 2001.
16. M. Talagrand. *The Generic Chaining. Upper and Lower Bounds for Stochastic Processes*. Springer, Berlin, 2005.